

## DOCUMENT RESUME

ED 346 710

FL 020 187

AUTHOR Milanovic, Michael; And Others  
TITLE Developing Rating Scales for CASE: Theoretical Concerns and Analyses.  
PUB DATE Feb 92  
NOTE 38p.; Paper presented at the Annual Language Testing Research Colloquium (14th, Vancouver, British Columbia, Canada, February 27-March 1, 1992). Appendix 2, "CASE Mnemonic Scales," will not reproduce well due to small print.  
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)  
EDRS PRICE MF01/PC02 Plus Postage.  
DESCRIPTORS Foreign Countries; \*Informal Assessment; Language Proficiency; Language Research; \*Language Tests; \*Rating Scales; Research Methodology; \*Test Construction  
IDENTIFIERS \*Cambridge Assessment of Spoken English

## ABSTRACT

The development of the Cambridge Assessment of Spoken English (CASE) is reviewed, with particular reference to the trialling and validation of the rating scales. The objectives of the test are to assess an individual's ability to produce and process spoken English; demonstrate the ability to communicate in English through the use of grammatical, discourse, and interactive skills; and maintain fluent and accurate discourse as appropriate. The test was based on a model of language proficiency. Attention was paid to the development process itself, which was both cyclical and iterative rather than linear. Decisions were evaluated in light of experience and trialling evidence and changes of direction applied as appropriate. Materials were developed in parallel with rating scales, training, and administrative procedure. The document includes approximately 17 pages of tabular or graphic information to supplement the narrative. Contains 7 references. (LB)

\*\*\*\*\*  
\* Reproductions supplied by EDRS are the best that can be made \*  
\* from the original document. \*  
\*\*\*\*\*

359

ED346770

## Developing Rating Scales for CASE: Theoretical Concerns and Analyses

Michael Milanovic  
Nick Saville  
Annette Cook

LTRC  
February 1992

PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

*Michael Milanovic*

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION  
Office of Educational Research and Improvement  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

☒ This document has been reproduced as  
received from the person or organization  
originating it.  
☐ Minor changes have been made to improve  
reproduction quality.

• Points of view or opinions stated in this docu-  
ment do not necessarily represent official  
OERI position or policy.

FL 020187

## **1. Introduction**

This paper reviews the development of the Cambridge Assessment of Spoken English (CASE), with particular reference to the trialling and validation of the rating scales. CASE was originally envisaged as an oral test for in-house company use. However, it soon became apparent that it could be used in a wider context. In December 1989, a working group established three objectives for the test. These were to produce a test that would assess an individual's ability to:

- i. produce and process spoken English;
- ii. demonstrate the ability to communicate in English through the use of grammatical, discourse and interactive skills;
- iii. maintain fluent and accurate discourse as appropriate.

From the inception of the project it was decided to base the test on a model of language proficiency. This was termed the model-based approach. Throughout the 1980's, testing theory has been moving steadily towards defining and validating models of communicative competence. This is demonstrated by the work of Canale and Swain (1980,1981), and most recently Bachman (1990, 1991). CASE is based primarily on the Bachman (1990, 1991) model.

Attention was also paid to the development process itself. There has been a tendency for test development to be linear in approach. From the outset, it proved impossible to follow a linear approach in the development of CASE, since this implies that, once decisions are made, they dictate future progress and cannot be readily changed in light of later findings. With the development of CASE it proved necessary to be both cyclical and iterative. Decisions were evaluated in the light of experience and trialling evidence and changes of direction applied as appropriate. Materials were developed in parallel with rating scales, training and administrative procedures.

## 2. The Structure of CASE

CASE consists of three phases. The first is a structured conversation, which is a 1-1 interaction between an interlocutor, normally a native speaker (NS), and a non-native speaker (NNS) candidate. While the interlocutor acts as an assessor, there is also a second, passive assessor present. In this phase, which is dominated primarily by the interlocutor, a fairly fixed range of topics, related to everyday life, is covered in a period of about 5 minutes. This phase is called the **interview**. The interlocutor is provided with a frame of topics and questions to be dealt with - the interlocutor frame. S/he is expected to follow this frame closely for all candidates, although, clearly the nature of the interaction is influenced by several factors such as background, personality and competence of the candidate. The range of topics covered in this phase include:

- greetings and introductions;
- giving information about self related to:
  - current job or status;
  - work and travel;
- a built in topic switch;
- future career prospects;
- interests;
- closing exchanges.

The second phase is called the **presentation**. In this phase, two candidates are asked to prepare a short presentation on a topic selected from a choice of three each. They are allowed about 5 minutes for the preparation phase. Each candidate then gives a 1-2 minute presentation. After each presentation, there is a short question and answer session, in which the candidate who is not giving the presentation is required to ask questions of the presenter. While there is some

NS involvement in this phase, the interaction is largely NNS/NNS. In this phase, the candidate is expected to:

- structure discourse;
- organise information;
- illustrate and exemplify;
- show awareness of an audience;
- respond to comments questions and objections;
- demonstrate linguistic control.

The third and final phase is called **discussion**. In this phase, candidates are engaged in an activity that involves the discussion<sup>4</sup> of a problem, negotiation and then arriving at consensus. The phase lasts 3-4 minutes. It is largely a NNS/NNS interaction although the scene is set by the NS interlocutor, who may intercede in the discussion if required. In this phase, the candidate is expected to:

- negotiate meaning;
- manage conversational direction;
- take turns;
- summarise own and partner's position;
- assess responses in real time;
- monitor and adjust linguistic performance.

The overall structure and timing of CASE is illustrated graphically in Appendix 1.

CASE differs in important ways from many other oral interviews. Firstly, it is a paired interview that allows for a variety of interactions - NS/NNS, NS/NNS-NNS, NNS/NNS. Secondly, candidates engage in a range of activities that allow for a broad sampling of overall ability. Thirdly, the pairings in phases two and

three are based on an initial impression of compatibility and ability arrived at in phase 1. Briefly, candidates take CASE in groups of six. After phase 1, which a series of six 1-1 interactions, the raters pair the candidates for phases two and three. As mentioned above, this pairing is based on an initial impression of compatibility and ability.

### **3. The Rating Scales**

The rating scales were developed along side the structure of CASE and in light of the model of communicative competence that is being assessed. Essentially the scales break down into three main areas. These are grammatical, discourse and strategic competence. There are three grammatical scales: grammar, vocabulary and pronunciation. Each of these has 5 points. Two scales are related to discourse competence: organisation and communication skills and strategies, although it must be said that the latter scale overlaps with strategic competence. These scales have five points. Four scales are related to strategic competence, and focus on the success with which the four major tasks in CASE are achieved. These main tasks are the interview, the presentation, the questions and the discussion. These scales have four points. Another scale, called interlocutor support, is also included. This was intended to quantify the amount of support a candidate might require from the interlocutor during CASE. This scale has four points. Unfortunately, data related to this scale was not gathered effectively.

While the scales described above attempt to focus on the components of communicative language ability (CLA), it was also decided to develop a scale related to overall competence. It must be pointed out that this scale is not a summation of the component scales. Rather it is an independent assessment of overall spoken language ability. The overall ability scale has six points. In conclusion, CASE employs eleven rating scales in all. 10 are related to components of CLA and the 11th is a measure of overall ability. The scales are

summarised in Appendix 2. It should be noted that the scales are of different lengths. When CASE was first developed not only were all the scales of the same length, they also covered nine points. It soon became apparent that raters were not able to differentiate effectively over all the scales. The scale lengths were therefore adjusted in light of feedback from raters. In addition, the overall scale was made substantially different from the component scales in order to avoid contamination of either scale by the other. This is well illustrated by reference to Appendix 2. The extent to which raters would be able to differentiate effectively using the revised scales was considered to be an empirical question. This will be discussed below.

#### **4. The Validation of the CASE Rating Scales**

Three major trialling exercises have taken place to date. The first in Hong Kong, the second in Thailand and the third in the UK. The Hong Kong and Thailand trials focused primarily on administrative feasibility and face validity. The question of rating scales was addressed, but only from the point of view of subjective feedback from raters and analysis of video recordings. This feedback suggested that the scales were too long and sometimes irrelevant. Corrective action was taken. The UK trials have focused on training activities and scale validation. These trials will be reported on here.

##### **4.1 The UK Trials**

Between January and September 1991 a total of 70 candidates were assessed in Norwich, Cambridge, Cardiff and Newcastle. They came from a variety of backgrounds but were mostly working in a business context or students. All interviews were either audio or video recorded. Approximately half the interviews were transcribed. In all 10 raters were trained and participated in the study. All of them came from an EFL background and were practising teachers at the time of the study. All received at least one full day of training and

standardisation before they engaged in rating. Four of them rated on more than one occasion, while the remaining six rated on one occasion only.

Each interview requires the presence of two raters, one acting as interlocutor, the other acting as an independent assessor. Five of the raters engaged in both roles at various times. Scores were recorded in specially prepared score booklets and then transferred to a database system developed using a programme called Crystal, which is a PC-based expert system. Rules were created that enabled the scores of the two assessors to be reconciled to create profile statements. The scores that the scale validation is based on have not been adjusted in any way.

#### **4.2 Inter Rater Reliability**

The first question addressed was the degree of inter rater agreement. Correlations between the first and second ratings on the eleven scales were computed using SPSSPC. The correlation between the overall ratings of first and second rater was established at 0.93. The inter correlations of the component scales varied between 0.47 and 0.90. Appendix 3 provides a complete correlation matrix of the scales across raters. The correlations are high for the most part. There are several possible reasons for this:

- i. phase 1 pairing serves as an initial assessment and allows the raters to discuss candidates prior to arriving at a final assessment;
- ii. an opportunity is supplied for the pairing in phases 2 and 3 to provide a conducive inter personal environment;
- iii. training and standardisation procedures are effective and rigorous;
- iv. the presence of two raters allows for an ongoing standardisation process;
- v. the length of the scales reflects the degree to which raters can make accurate decisions;
- vi. the procedure is actually highly reliable;

- vii. the range and nature of the language sampled can be assessed using the scales.

### **4.3 Scales and Construct**

Initial investigation of the extent to which the scales reflect the model of CLA upon which they are based has employed correlational analysis. The analyses focused both within rater and between raters.

#### **4.3.1 Between Rater Analysis**

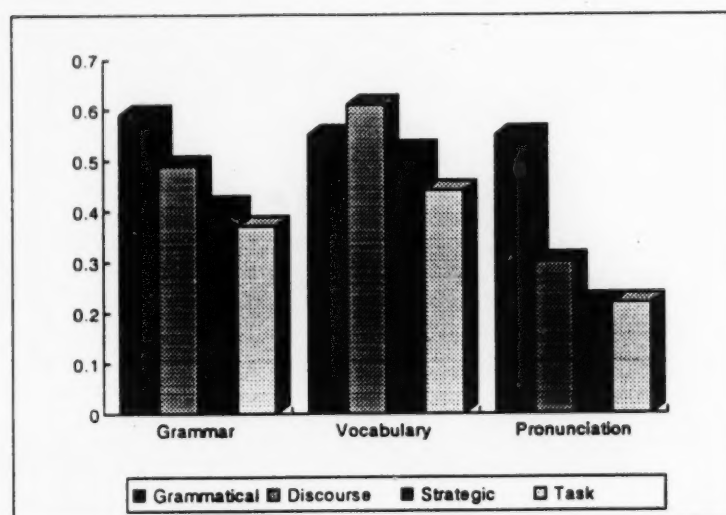
The scales were based on three basic elements of CLA - grammatical, discourse and strategic and a series of four task achievement scales. In order to investigate the extent to which the scales associated with these elements converged mean inter rater correlations were computed for the various scales. For example, the mean inter correlation of grammar on the first rating with the three components of grammatical competence in the second rating was established at 0.77. See Appendix 3. The inter correlation of grammar with discourse competence in the second rating was then established at 0.70, the inter correlation with strategic competence at 0.65 and with the tasks at 0.61. The same procedure was conducted for each of the scales in turn. A Fisher Z transformation was not employed since it makes virtually no difference with mid-range correlations. In order to look at the amount of shared variance,  $r^2$ 's were then calculated.

Table 1 below shows the  $r^2$ 's of the grammar, vocabulary and pronunciation scales with the elements of CLA and the task scales.

**Table 1: Inter rater  $r^2$  and components of CLA**

|             | Grammar | Vocabulary | Pronunciation |
|-------------|---------|------------|---------------|
| Grammatical | .59     | .55        | .55           |
| Discourse   | .49     | .61        | .30           |
| Strategic   | .41     | .52        | .22           |
| Task        | .37     | .44        | .22           |

**Graph 1**



The vocabulary scale converges more with discourse competence than it does with grammatical competence. This seems to make sense. Graph 1 illustrates the relationships very clearly.

It is interesting the extent to which pronunciation relates better to grammatical competence than it does to either strategic, discourse or task achievement. However, this is not surprising. Pronunciation does not need to be closely

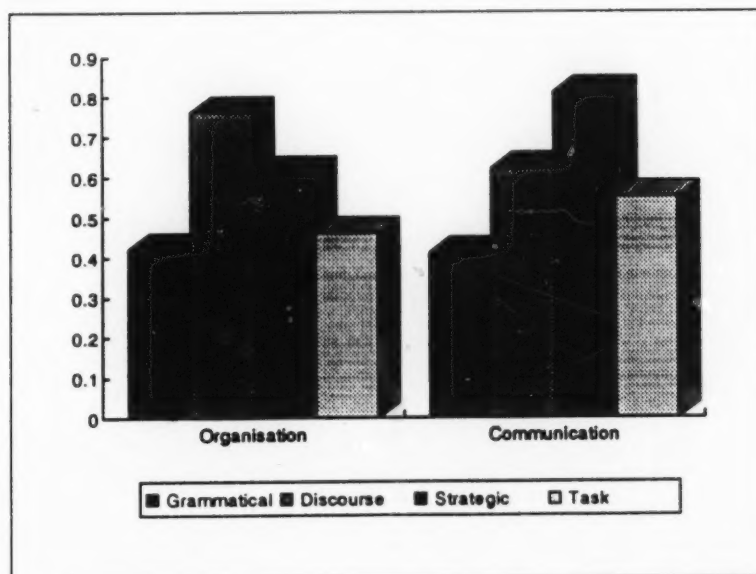
associated with an individual's ability to function effectively in a foreign language.

Table 2 illustrates the relationship between discourse and strategic competence, the components of CLA and task achievement. There appears to be a stronger relationship between discourse and strategic competence than between grammatical and strategic.

**Table 2: Inter rater  $r^2$  and components of CLA**

|             | Organisation | Communication |
|-------------|--------------|---------------|
| Grammatical | .42          | .41           |
| Discourse   | .76          | .62           |
| Strategic   | .61          | .81           |
| Task        | .46          | .55           |

**Graph 2**



Grammatical competence relates in much the same way to both the organisation and communication scales. However, the communication scale has a stronger relationship with task achievement than does organisation.

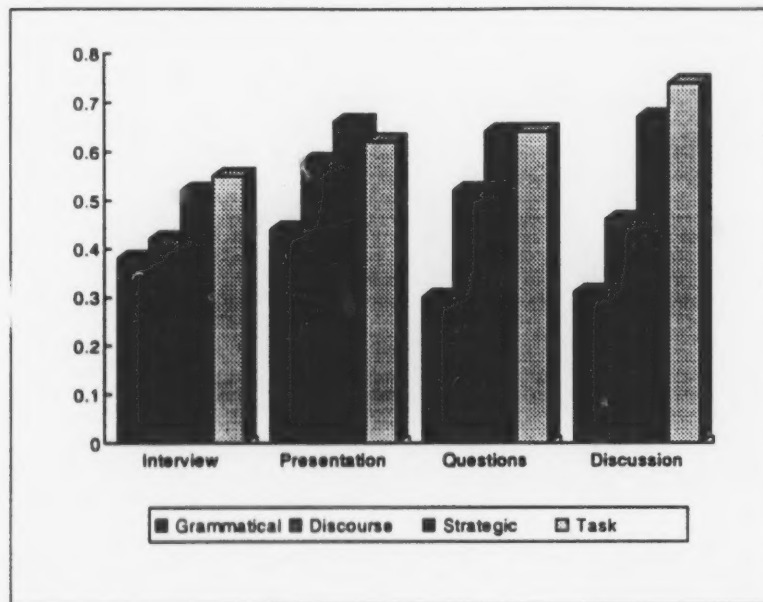
Table 3 below illustrates the relationship between the four task related scales and the components of CLA. In all cases the weakest relationship is with grammatical competence particularly with the questions and discussion. These are the most spontaneous of the of the tasks and this may well be a reason for the weaker relationship.

**Table 3: Inter rater  $r^2$  and components of CLA**

|             | Interview | Presentation | Questions | Discussion |
|-------------|-----------|--------------|-----------|------------|
| Grammatical | .38       | .44          | .30       | .31        |
| Discourse   | .42       | .58          | .52       | .46        |
| Strategic   | .52       | .66          | .64       | .67        |
| Task        | .55       | .62          | .64       | .74        |

The communication scale, which represents strategic competence, consistently has a stronger relationship with the achievement of tasks than does the organisation scale. In the presentation task, the organisation scale correlates more highly than it does elsewhere. This finding is in line with intuition. It makes sense that organisation should be best evaluated during the presentation. Indeed, the presentation task was designed to allow for a view of discourse skills to be best achieved.

**Graph 3**



While inter scale correlations are consistently high, which might indicate some sort of halo effect, it is also apparent that raters are able to distinguish between the various scales.

#### 4.4 The Integrity of the Scales

There are many difficulties associated with the construction of rating scales. For example, there is the issue of what to describe. In this project, the nature of the scales was dictated to a large extent by the model of CLA employed. One way of validating the integrity of the scales in this sense is through some form of correlational analysis. This was attempted in the previous section. Once what to describe has been established, the degree to which raters are able to differentiate between the points on the scale needs to be investigated. As mentioned above, this can be done through feedback from raters. This method was employed in the first instance after the trials in Hong Kong and Thailand.

Partial credit analysis (Wright and Masters, 1982) provides a means for the empirical validation of rating scales in a way that feedback from raters does not. The partial credit model allows for the scoring of test items in any number of ordered categories and is an appropriate analysis tool for rating scales (Adams et al., 1987, Pollitt and Hutchinson, 1987). The model describes the relationship between a person's ability and the difficulty of the task or scale in terms of the probability of an individual providing an adequate language sample to get a certain score.

For the purposes of this study, the rating scales were analysed in three ways using BIGSTEPS (Wright and Linacre, 1991): the ratings for rater 1, the ratings for rater 2 and the ratings for both raters. The results are most clearly illustrated through the item characteristic curves, which will be discussed below. However, it is first necessary to look at the degree to which the model fits the data. Table 4 below is a measure of person fit. In this case the table refers to the first rater. The results are similar for the second rater and both raters. The results of these analyses are not discussed here.

Several points can be noted from this table. First, the mean score of the persons in the study across all the scales is reported in the 'SCORE' column. The score is 35.6 (60%) out of a possible total score of 59. The standard deviation is reported just below at 7.9 (13%). The 'COUNT' column refers to the number of items or in this case scales analysed.

**Table 4: Summary of 68 measured persons**

|                                              |            |          |           |                          |          |                 |             |
|----------------------------------------------|------------|----------|-----------|--------------------------|----------|-----------------|-------------|
| CASE Analysis : Feb 1992                     |            |          |           | R MODEL "BIGSTEPS" RASCH |          |                 |             |
| INPUT:                                       | 70 PERSONS | 11 ITEMS | ANALYSED: | 68 PERSONS               | 11 ITEMS | 47 CATEGORIES   |             |
| SUMMARY OF 68 MEASURED (NON-EXTREME) PERSONS |            |          |           |                          |          |                 |             |
|                                              | SCORE      | COUNT    | MEASURE   | ERROR                    | MNSQ     | INFIT           | MNSQ OUTFIT |
| MEAN                                         | 35.6       | 11.0     | .11       | .64                      | .98      | -.2             | .98         |
| S.D.                                         | 7.9        | .0       | 3.26      | .12                      | .59      | 1.2             | .71         |
| RMSE                                         | .65        | ADJ.S.D. | 3.19      | PERSON SEP               | 4.90     | PERSON SEP REL. | .96         |
| MAXIMUM EXTREME SCORE:                       |            |          |           | 2 PERSONS                |          |                 |             |

The 'MEASURE' column refers to the average ability of the group of candidates in logits, 0.11 and the ability spread or standard deviation of the group in logits, 3.26. The spread of ability is quite broad. The 'ERROR' column reports the mean error in logits, 0.64, and its standard deviation, 0.12. The weighted fit statistics are reported in the 'MNSQ' (mean square) and 'INFIT' columns. Perfect fit would result in a 'MEAN' of 0 and a 'SD' of 1 in the 'INFIT' column. As can be noted from the table the data fit the model well. The 'MNSQ' and 'OUTFIT' columns are unweighted measures. There is a slight degree of misfit, though this is not significant. The 'RMSE' (root mean standard error) measure is equivalent to a traditional standard error of measurement and the 'ADJ.S.D.' (adjusted standard deviation) is the standard deviation corrected for measurement error. The 'PERSON SEP' measure shows how many times wider

the ability scale is than the error in that scale. It is calculated as a ratio of ADJ.S.D. over RMQS ( $3.19/0.65$ ). The 'PERSON SEP REL' measure is an indication of the reliability of the test. It is similar to a coefficient alpha. In all we can be fairly satisfied with the scales in this data.

Table 5 is essentially a representation of the quality of measurement. The two horizontal dotted lines at 2 and -2 are at the 5% significance band. The numbers on the table represent people.

Table 5: Person Fit

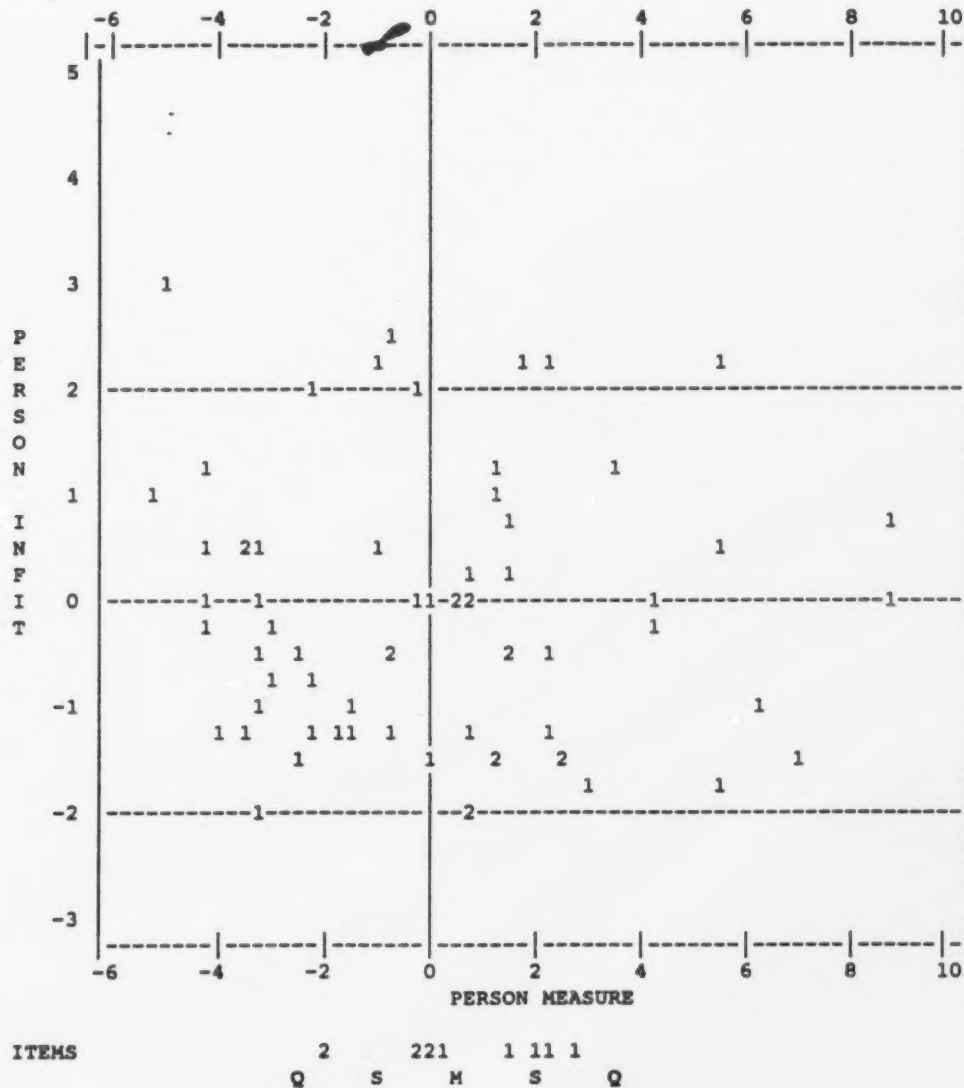
CASE Analysis : Feb 1992

MODEL "BIGSTEPS" RASCH

INPUT: 70 PERSONS 11 ITEMS

ANALYSED: 68 PERSONS 11 ITEMS

47 CATEGORIES



It will be noted that nobody misfits at the bottom. Misfit at this end might well indicate a flat profile and suggest that a halo effect were coming into play. This is not the case. There is some slight misfit at the top. This is likely to be due to a

particularly jagged profile. It will be noted that there is only a limited amount of misfit. From this table we are able to conclude that the people fit the model well.

Table 6 is a representation of the raters' sensitivity to the scales. For ease of reference the 3 point on the scales have been changed to an asterisk. The table shows where raters are prepared to give credit for ability or not as the case may be. For example it can be noted that a person most is likely to get a band 2 on the grammar and vocabulary scales at any ability level of -3 logits. It appears easier to get a band 3 on the pronunciation scale, and the length of the 3 band in the pronunciation scale is rather long. This may indicate that raters are hedging their bets to some extent when it comes to pronunciation. This is not a counter intuitive finding, however. We all know from experience that there is a tendency for pronunciation to fossilise. Perhaps the scale is a representation of this tendency. Interestingly, the bottom point on the grammar, vocabulary and pronunciation scales does not come into play. This may be an indication that the scales are too long, and that raters cannot differentiate at five points. It may mean that no one was at the one point in terms of these particular scales. Or it may mean that the scale descriptions need clarification at the lowest ability levels. In any case, this question will be addressed as the validation of the scales continues.

The bottom point on the scales is used in the organisation and communication scales. This finding is a further indication that raters are able to differentiate between the grammatical and discourse components of CLA. Effective organisation skills are as much a feature of NS discourse as they are of NNS discourse.



In the validation of rating scales it is also important to establish the extent to which the points on a scale can be effectively differentiated by raters. A very graphic way of doing this is to study the ICC tables for each of the scales (see Appendix 4). ICCs show how the modelled probability of getting a particular point on the scale varies with the ability of the people in question. The first table in Appendix 4 illustrates the overall ability scale. The five points, from 3 - 7 are fairly clearly differentiated. There is always a range of ability where a particular scale point is the most probable score. The degree of overlap between the five points is quite limited and the overlap is rarely more than one point on the scale. Points 4 and 5 on the scale are the least discrete, while points 3, 6 and 7 are reasonably distinct. Point 6 seems to have the greatest spread, however, the probability of an individual getting a 4 instead of a 6 is low.

The peakedness of the curves is also an important feature of ICCs. The greater the peakedness, the higher the probability that an individual with an ability in line with the peak is being correctly assessed. In conjunction with the peakedness, it is important to consider the spread of the ICCs. The wider the spread, the greater the possibility of inappropriate scoring. The three ICCs related to grammar, vocabulary and pronunciation seem to indicate that the scale points are well differentiated. However, the organisation and communication scales show that the probability of achieving the second point on the scale is relatively low. This may be due to a lack of differentiation in the scale descriptor, it may indicate that a five point scale is too long, or it may suggest that the abilities measured by these two scales have a rather strange step action. By this we mean that the abilities may be either not present (1) or present to some reasonable degree (3) with little in between. The task achievement scales are all similar to the extent that the probability of scoring the mid points is lower than that of scoring the extreme points it may be that the descriptors for the mid points need some work, or that raters find it more straightforward to judge good

and bad performances than they do mediocre. In any case, some attention to these scales will be required in light of this validation study and prior to further trialling.

## **5. Conclusion**

The principled development and validation of the CASE rating scales has proven to be a valuable exercise. It has supplied the test developers with many useful insights and ammunition for continued work. The next stage of validation will be to establish the relative harshness of the raters and scales. This will be done through multi-faceted Rasch analysis. In addition, it will be necessary to carry out a detailed analysis of the intrinsic difficulty of the tasks themselves.

## References

- Adams, R.J., P.E. Griffin and L. Martin, 1987, A latent trait method for measuring a dimension in second language proficiency, *Language Testing*, Vol. 4, No. 1.
- Canale, M., and M. Swain, 1980, Theoretical bases of communicative approaches to second language teaching and testing, *Applied Linguistics* 1, 1-47.
- Bachman, L.F., 1990, *Fundamental Considerations in Language Testing*, Oxford, OUP.
- Bachman, L.F., 1991 (forthcoming), Manuscript
- Pollitt, A. and Hutchinson, C., 1987, Calibrating graded assessments: Rasch partial credit analysis of performance in writing, *Language Testing*, Vol. 4, No. 1.
- Wright, B.D., and M. Linacre, 1991, *A User's Guide to Bigsteps*, Chicago, Mesa Press.
- Wright, B.D., and G.N. Masters, 1982, *Rating Scale Analysis: Rasch Measurement*, Chicago, Mesa Press.

# Appendix 1: The Structure of CASE

|                  |                             |                    |                     |                      |
|------------------|-----------------------------|--------------------|---------------------|----------------------|
| 5 Mins.          | 5 Mins.                     | 1 Min.             | 8 Mins.             | 4 Mins.              |
| <b>Stage One</b> |                             | <b>Stage Two</b>   |                     |                      |
|                  |                             | <b>Section one</b> | <b>Section Two</b>  | <b>Section three</b> |
| <b>Interview</b> | <b>Deciding on Pairings</b> | <b>Warm-up</b>     | <b>Presentation</b> | <b>Discussion</b>    |

| <b>Case Scales</b> |                                                                                                           |                                                                           |                                                                                              |                                                                                                             |                                                                                             |                             |                                                               |
|--------------------|-----------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|-----------------------------|---------------------------------------------------------------|
| <b>Rating</b>      | <b>Grammar</b>                                                                                            | <b>Vocabulary</b>                                                         | <b>Pronunciation</b>                                                                         | <b>Organisation</b>                                                                                         | <b>Communication Strategies and Interaction</b>                                             | <b>Interlocutor Support</b> | <b>Task Achievement</b>                                       |
| <b>A</b>           | Extensive range structures including complex used accurately; very few basic errors.                      | Extensive range; accurate and appropriate; wide topic range.              | Accurate and consistent use all aspects of pron.                                             | Excellent - well developed logical explanation.                                                             | Initiates and maintains interaction; responds with ease to topic shifts.                    | Not required                | Task fully achieved and communicated successfully.            |
| <b>B</b>           | Large range basic structures used with few errors; errors with complex structures.                        | Large range of everyday vocab accurately used.                            | Broadly accurate and consistent use most aspects pron.; foreign influence doesn't interfere. | Good - clear explanations with extended discourse well organised; most discourse relationships well marked. | Intended meaning communicated in most contexts; initiates and maintains interaction.        | Occasionally required       | Task is achieved but 1 or more of task requirements lacking.  |
| <b>C</b>           | Moderate range of basic structures accurate; occasional accurate use complex structures.                  | Moderate range for everyday use.                                          | Often faulty but intelligible; foreign accent doesn't obscure intelligibility.               | Moderate effectiveness with extended discourse; need for some clarification and explanation.                | Communicates main ideas; Uses repair strategies; difficulties in initiating.                | Frequently required         | Task only partly achieved; several task requirements lacking. |
| <b>D</b>           | Restricted range of structures; many inaccuracies in basic structures.                                    | Restricted range sufficient for basic communication only; choice limited. | Frequent errors causes some unintelligibility.                                               | Restricted - extended discourse; badly organised; only basic conversational routines.                       | Main idea communicated in limited contexts; repair strategies rarely used; initiation rare. | Continually required        | Task not achieved; task requirements not fulfilled.           |
| <b>E</b>           | Very restricted range; inaccurate use of basic structures; no complex structures; largely unintelligible. | Very restricted range usually inadequate for clear communication.         | Largely unintelligible.                                                                      | No evidence of extended discourse or conversational rules.                                                  | Great difficulty in communicating; unable to use repair strategies.                         |                             |                                                               |

|                        |                              |                         |                              |                                |                                    |                                   |
|------------------------|------------------------------|-------------------------|------------------------------|--------------------------------|------------------------------------|-----------------------------------|
| <b>Overall ability</b> | <b>800</b><br>Very Good User | <b>700</b><br>Good User | <b>600</b><br>Competent User | <b>500</b><br>Independent User | <b>400</b><br>Threshold level User | <b>300</b><br>Waystage Level User |
|------------------------|------------------------------|-------------------------|------------------------------|--------------------------------|------------------------------------|-----------------------------------|

BEST COPY AVAILABLE

| Correlations: Overall |        | GRAM1  | VOCAB1 | PRON1  | ORG1   | COMM1  | SUPP1  | INT1   | PRES1  | QUES1  | DISC1  |
|-----------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| TOT1                  | 1.0000 | .7352  | .7524  | .6413  | .7262  | .7461  | .5817  | .7170  | .7569  | .7232  | .7240  |
| GRAM1                 | .7352  | 1.0000 | .8277  | .7142  | .7382  | .7252  | .6027  | .6931  | .6909  | .6095  | .6107  |
| VOCAB1                | .7524  | .8277  | 1.0000 | .6401  | .7557  | .7191  | .6099  | .6802  | .7334  | .6623  | .6420  |
| PRON1                 | .6413  | .7142  | .6401  | 1.0000 | .5829  | .5779  | .5419  | .4706  | .6232  | .5028  | .4670  |
| ORG1                  | .7262  | .7382  | .7557  | .5829  | 1.0000 | .8142  | .4425  | .7058  | .7573  | .7485  | .6976  |
| COMM1                 | .7461  | .7252  | .7191  | .5779  | .8142  | 1.0000 | .4476  | .7490  | .7999  | .7634  | .7939  |
| SUPP1                 | .5817  | .6027  | .6099  | .5419  | .4425  | .4476  | 1.0000 | .6186  | .4626  | .4090  | .3945  |
| INT1                  | .7170  | .6931  | .6802  | .4706  | .7058  | .7490  | .6186  | 1.0000 | .7486  | .7348  | .7879  |
| PRES1                 | .7569  | .6909  | .7334  | .6232  | .7573  | .7999  | .4626  | .7486  | 1.0000 | .8246  | .8563  |
| QUES1                 | .7232  | .6095  | .6623  | .5028  | .7485  | .7634  | .4090  | .7348  | .8246  | 1.0000 | .8756  |
| DISC1                 | .7240  | .6107  | .6420  | .4670  | .6976  | .7939  | .3945  | .7879  | .8563  | .8756  | 1.0000 |
| TOT2                  | .9324  | .7064  | .7312  | .6125  | .7317  | .7502  | .6066  | .7175  | .7373  | .7174  | .7391  |
| GRAM2                 | .7208  | .8765  | .7895  | .7142  | .6819  | .7084  | .6457  | .7341  | .6708  | .5739  | .5925  |
| VOCAB2                | .7267  | .7441  | .8285  | .6541  | .6986  | .6378  | .5625  | .6450  | .7058  | .5843  | .5958  |
| PRON2                 | .6745  | .6850  | .6045  | .7630  | .5940  | .5714  | .5208  | .4760  | .5904  | .4911  | .4965  |
| ORG2                  | .7907  | .6967  | .7783  | .5496  | .8665  | .7858  | .4895  | .6590  | .7598  | .7166  | .6785  |
| COMM2                 | .7457  | .6482  | .7200  | .4738  | .7847  | .9049  | .4621  | .7177  | .8142  | .8048  | .8179  |
| SUPP2                 | .      | .      | .      | .      | .      | .      | .      | .      | .      | .      | .      |
| INT2                  | .7282  | .6029  | .6615  | .4273  | .6500  | .7222  | .4925  | .8098  | .7222  | .7705  | .8471  |
| PRES2                 | .7444  | .6898  | .7196  | .5576  | .7072  | .7618  | .4672  | .7379  | .8756  | .7499  | .8253  |
| QUES2                 | .7190  | .5606  | .6616  | .4630  | .7145  | .7315  | .4200  | .7082  | .7949  | .8937  | .8605  |
| DISC2                 | .6840  | .5818  | .5957  | .4223  | .6332  | .7310  | .4086  | .7041  | .7827  | .7924  | .8881  |

N of cases: 70      1-tailed Signif: \* - .01      - .001      70      1-tailed Signif: \* - .01      - .001

| Correlations: Overall 2 |        | GRAM2  | VOCAB2 | PRON2  | ORG2   | COMM2  | SUPP2  | INT2   | PRES2  | QUES2  | DISC2  |
|-------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| TOT1                    | .9324  | .7208  | .7267  | .6745  | .7907  | .7457  | .      | .7282  | .7444  | .7190  | .6840  |
| GRAM1                   | .7064  | .8765  | .7441  | .6850  | .6967  | .6482  | .      | .6029  | .6898  | .5606  | .5818  |
| VOCAB1                  | .7312  | .7895  | .8285  | .6045  | .7783  | .7200  | .      | .6615  | .7196  | .6616  | .5957  |
| PRON1                   | .6125  | .7142  | .6541  | .7630  | .5496  | .4738  | .      | .4273  | .5576  | .4630  | .4223  |
| ORG1                    | .7317  | .6819  | .6986  | .5940  | .8665  | .7847  | .      | .6500  | .7072  | .7145  | .6332  |
| COMM1                   | .7502  | .7084  | .6378  | .5714  | .7858  | .9049  | .      | .7222  | .7618  | .7315  | .7310  |
| SUPP1                   | .6066  | .6457  | .5625  | .5208  | .4895  | .4621  | .      | .4925  | .4672  | .4200  | .4086  |
| INT1                    | .7175  | .7341  | .6450  | .4760  | .6590  | .7177  | .      | .8098  | .7379  | .7082  | .7041  |
| PRES1                   | .7373  | .6708  | .7058  | .5904  | .7598  | .8142  | .      | .7222  | .8756  | .7949  | .7827  |
| QUES1                   | .7174  | .5739  | .5843  | .4911  | .7166  | .8048  | .      | .7705  | .7499  | .8937  | .7924  |
| DISC1                   | .7391  | .5925  | .5958  | .4965  | .6785  | .8179  | .      | .8471  | .8253  | .8605  | .8881  |
| TOT2                    | 1.0000 | .7208  | .7474  | .7183  | .7918  | .7811  | .      | .7461  | .7744  | .7234  | .6967  |
| GRAM2                   | .7208  | 1.0000 | .7652  | .6564  | .6410  | .6482  | .      | .6671  | .6522  | .5782  | .5643  |
| VOCAB2                  | .7474  | .7652  | 1.0000 | .6666  | .7442  | .6555  | .      | .6199  | .6971  | .6032  | .5317  |
| PRON2                   | .7183  | .6564  | .6666  | 1.0000 | .6574  | .6148  | .      | .4559  | .6251  | .5435  | .5757  |
| ORG2                    | .7918  | .6410  | .7442  | .6574  | 1.0000 | .7988  | .      | .6237  | .7396  | .7291  | .6559  |
| COMM2                   | .7811  | .6482  | .6555  | .6148  | .7988  | 1.0000 | .      | .7615  | .8027  | .8429  | .8194  |
| SUPP2                   | .      | .      | .      | .      | .      | .      | 1.0000 | .      | .      | .      | .      |
| INT2                    | .7461  | .6671  | .6199  | .4559  | .6237  | .7615  | .      | 1.0000 | .7561  | .8352  | .8079  |
| PRES2                   | .7744  | .6522  | .6971  | .6251  | .7396  | .8027  | .      | .7561  | 1.0000 | .7848  | .8284  |
| QUES2                   | .7234  | .5782  | .6032  | .5435  | .7291  | .8429  | .      | .8352  | .7848  | 1.0000 | .8724  |
| DISC2                   | .6967  | .5643  | .5317  | .5757  | .6559  | .8194  | .      | .8079  | .8284  | .8724  | 1.0000 |

N of cases: 70      1-tailed Signif: \* - .01      - .001      70      1-tailed Signif: \* - .01      - .001

Appendix 4

TABLE 21.1 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

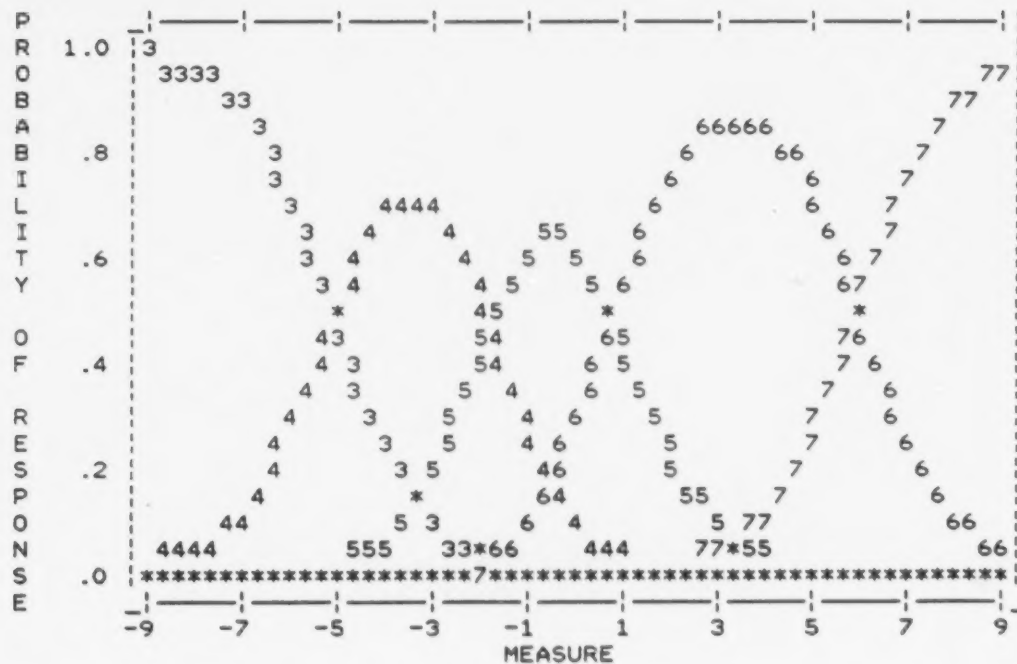
INPUT: 70 PERSONS 11 ITEMS

ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

PLOT CENTERED ON ITEM DIFFICULTY OF 1.14

RESPONSE CATEGORY PROBABILITY CURVES

OVERALL



# Appendix 4

TABLE 21.2 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

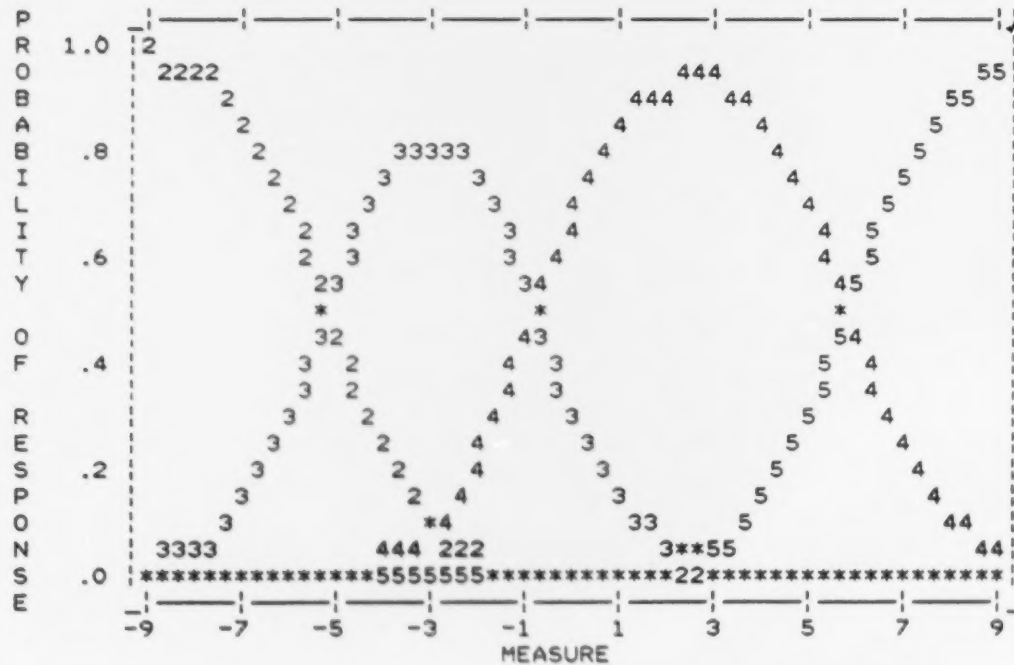
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 2

PLOT CENTERED ON ITEM DIFFICULTY OF 2.16

RESPONSE CATEGORY PROBABILITY CURVES

GRAMMAR



# Appendix 4

TABLE 21.3 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

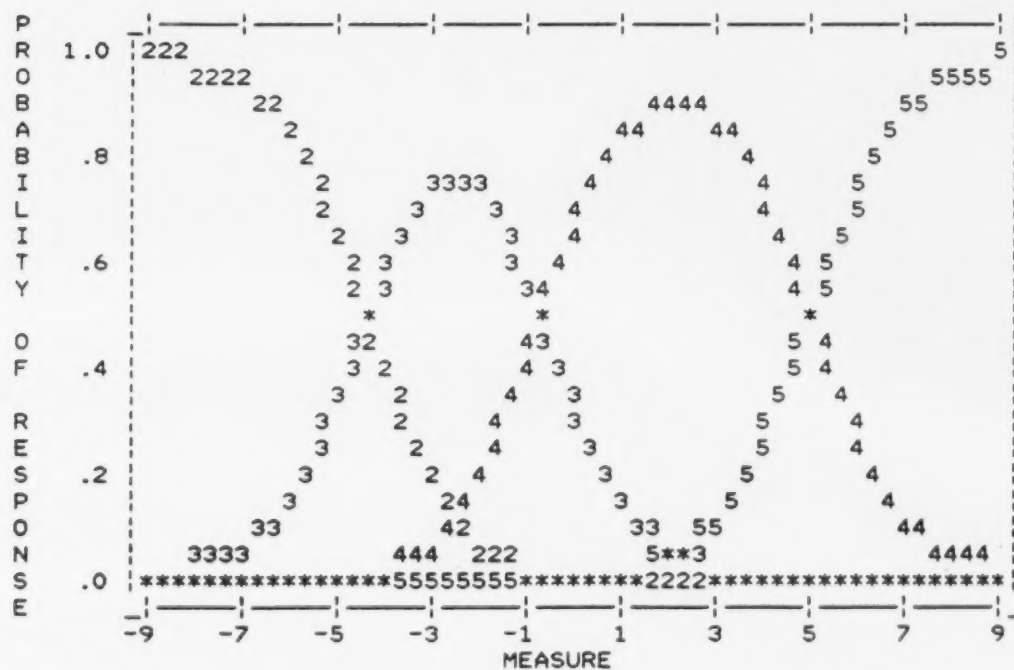
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 3

PLOT CENTERED ON ITEM DIFFICULTY OF 1.43

RESPONSE CATEGORY PROBABILITY CURVES

## VOCABULARY



Appendix 4

TABLE 21.4 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

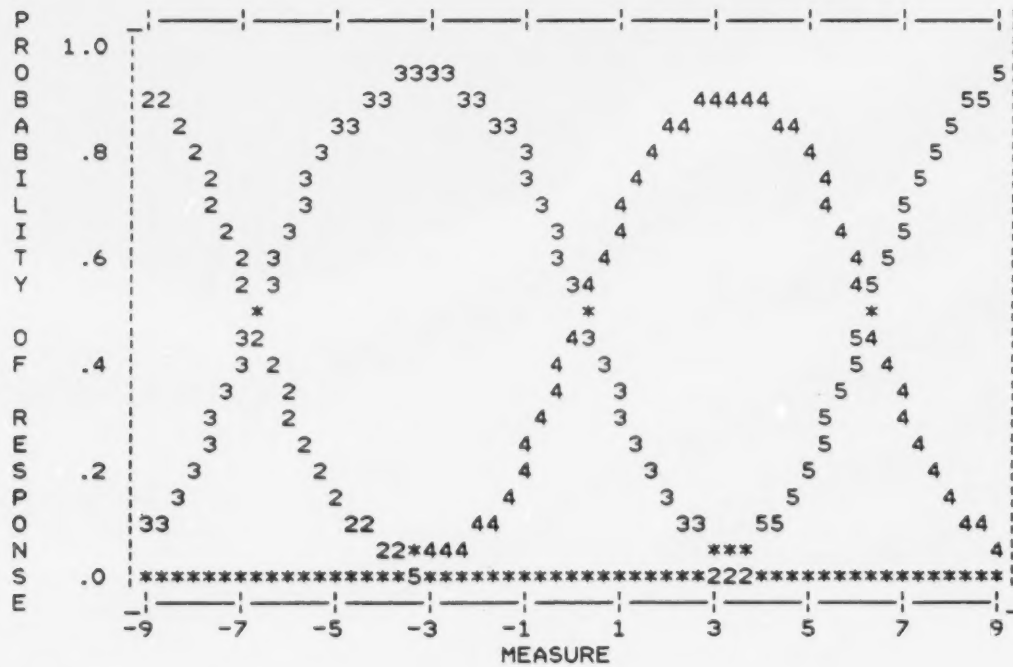
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 4

PLOT CENTERED ON ITEM DIFFICULTY OF 1.67

RESPONSE CATEGORY PROBABILITY CURVES

PRONUNCIATION



# Appendix 4

TABLE 21.5 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

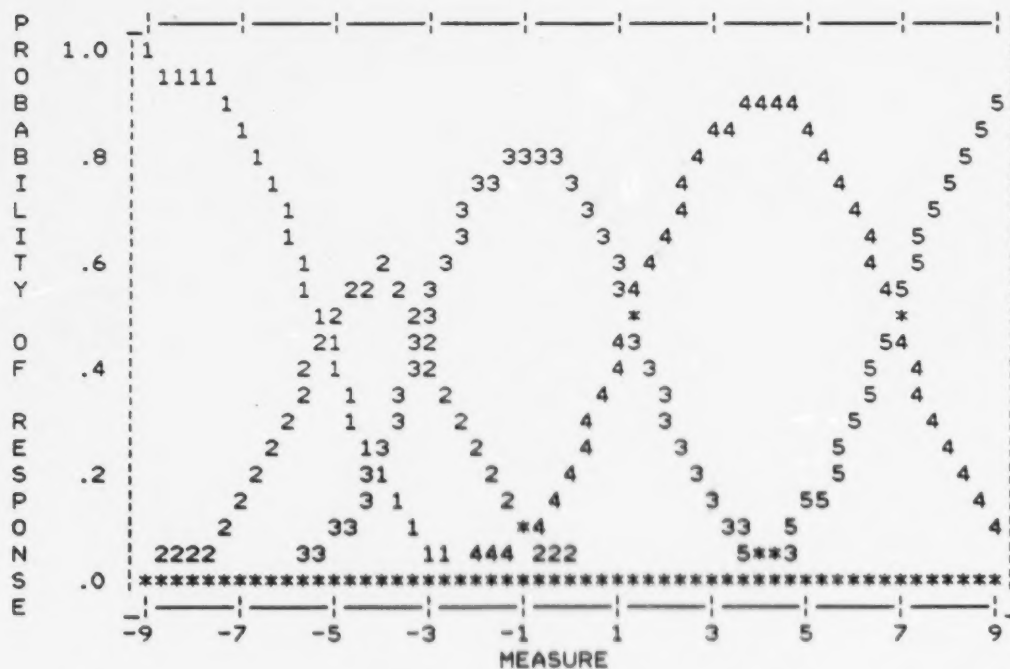
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 5

PLOT CENTERED ON ITEM DIFFICULTY OF -.38

RESPONSE CATEGORY PROBABILITY CURVES

ORGANIZATION



# Appendix 4

TABLE 21.6 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

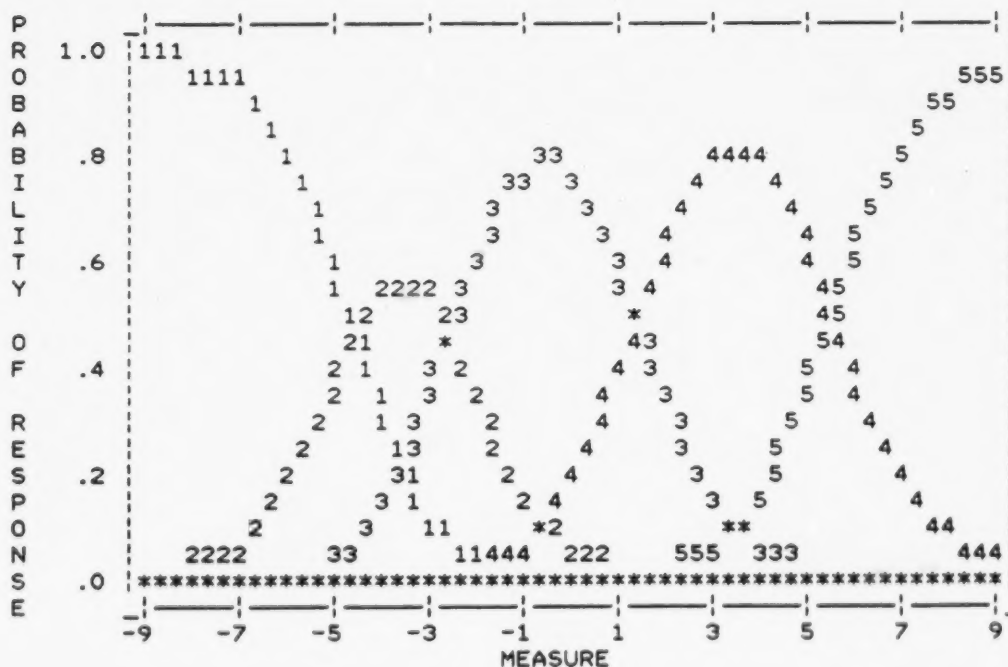
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 6

PLOT CENTERED ON ITEM DIFFICULTY OF -.61

RESPONSE CATEGORY PROBABILITY CURVES

COMMUNICATION STRATEGIES



# Appendix 4

TABLE 21.7 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

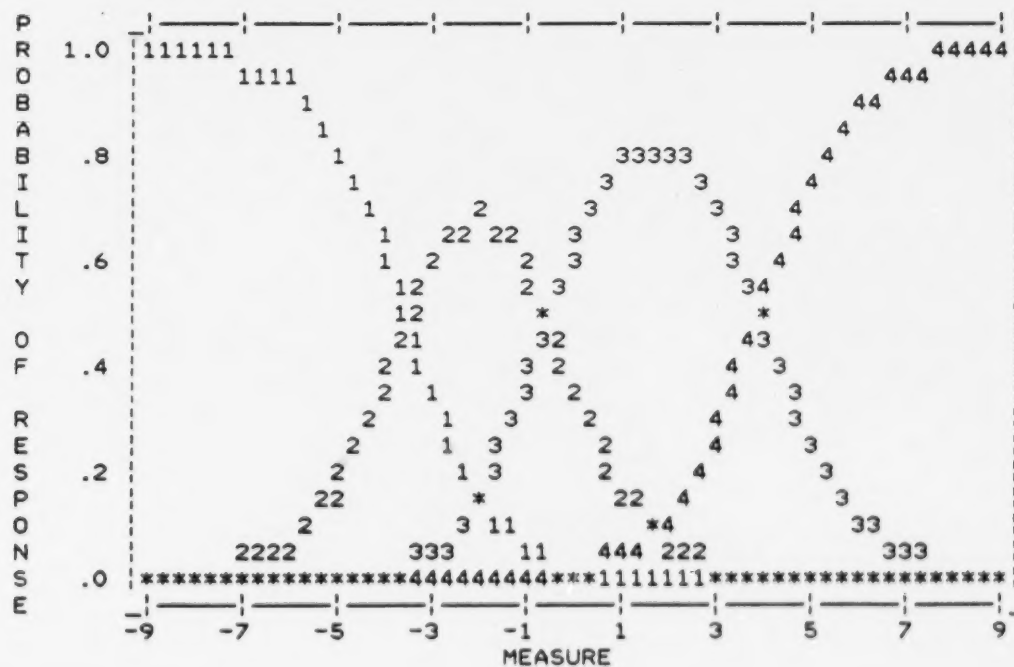
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 7

PLOT CENTERED ON ITEM DIFFICULTY OF -2.22

RESPONSE CATEGORY PROBABILITY CURVES

INTERLOCUTOR SUPPORT



# Appendix 4

TABLE 21.9 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

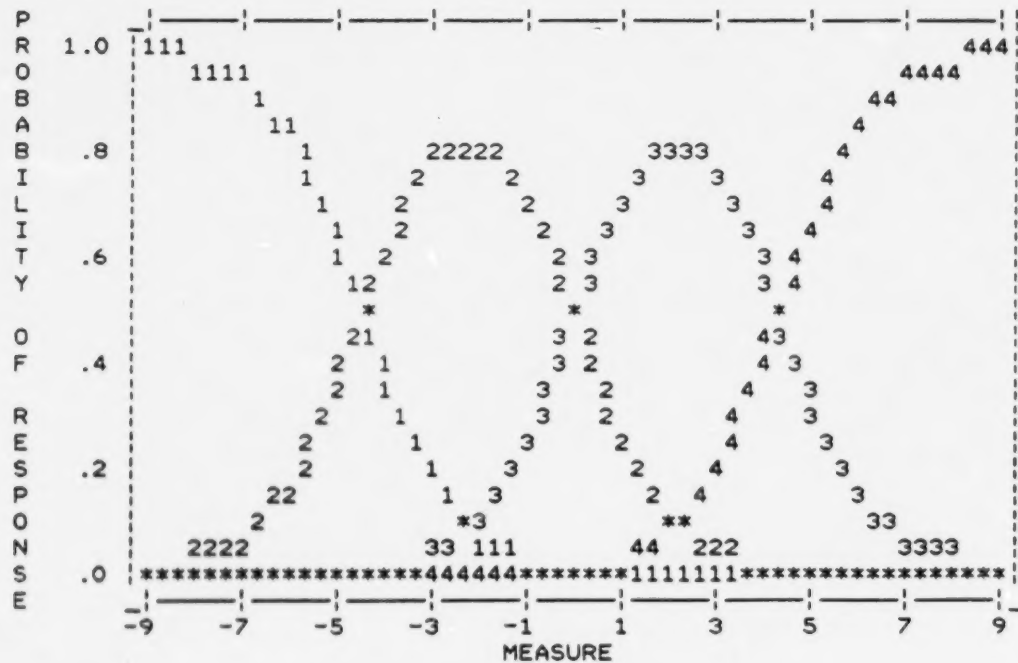
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 9

PLOT CENTERED ON ITEM DIFFICULTY OF -.16

RESPONSE CATEGORY PROBABILITY CURVES

## PRESENTATION



# Appendix 4

TABLE 21.10 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

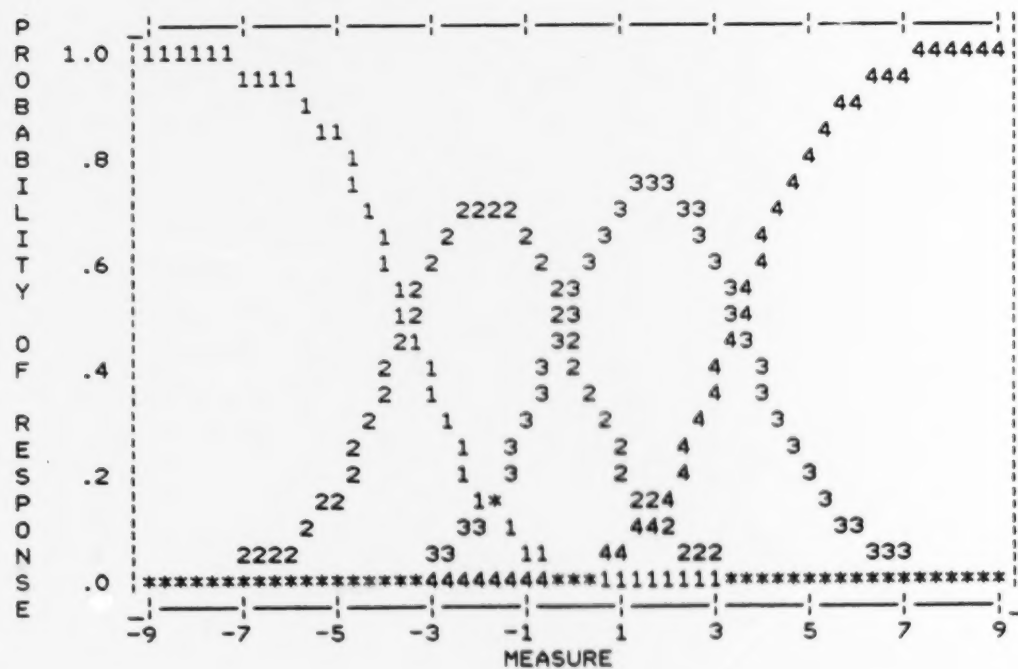
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 10

PLOT CENTERED ON ITEM DIFFICULTY OF -.26

RESPONSE CATEGORY PROBABILITY CURVES

## QUESTIONS



# Appendix 4

TABLE 21.11 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

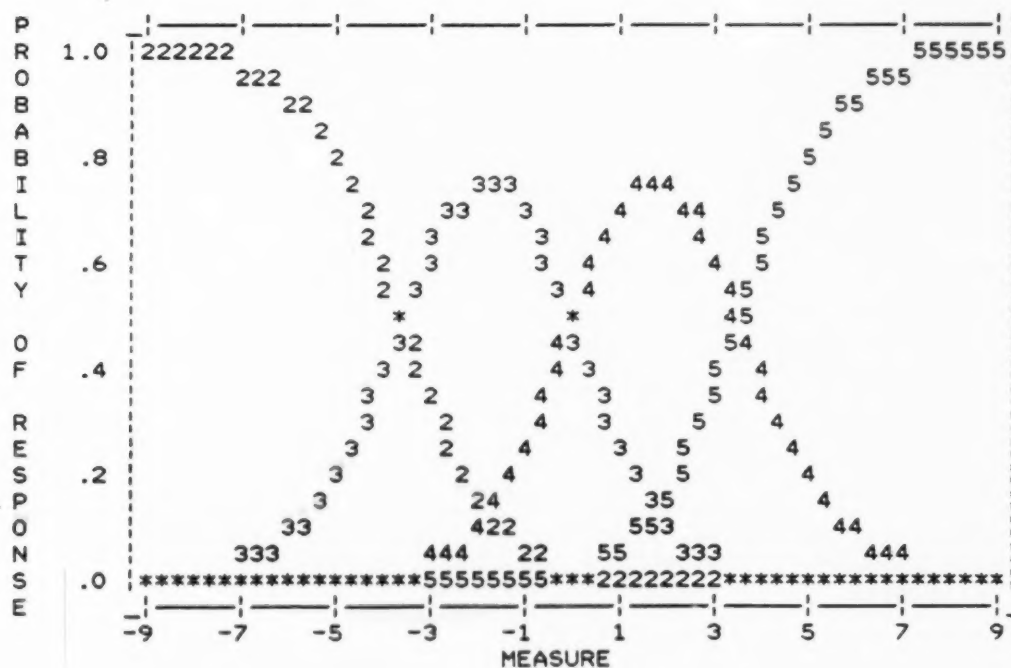
ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 11

PLOT CENTERED ON ITEM DIFFICULTY OF -.51

RESPONSE CATEGORY PROBABILITY CURVES

## DISCUSSION



Appendix 4

TABLE 21.8 CASE Analysis : Feb 1992

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12

Feb 20 16:56 1992

INPUT: 70 PERSONS 11 ITEMS

ANALYZED: 68 PERSONS 11 ITEMS 47 CATEGORIES

FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 8

PLOT CENTERED ON ITEM DIFFICULTY OF -2.26

RESPONSE CATEGORY PROBABILITY CURVES

INTERVIEW

